Artificial Intelligence in Ophthalmology: A Comparative Analysis of GPT-3.5, GPT-4, and Human Expertise in Answering StatPearls Questions

Year: 2023 | Citations: 110 | Authors: M. Moshirfar, Amal W Altaf, Isabella M. Stoakes, Jared J Tuttle, P. Hoopes

Abstract

Importance Chat Generative Pre-Trained Transformer (ChatGPT) has shown promising performance in various fields, including medicine, business, and law, but its accuracy in specialty-specific medical questions, particularly in ophthalmology, is still uncertain. Purpose This study evaluates the performance of two ChatGPT models (GPT-3.5 and GPT-4) and human professionals in answering ophthalmology questions from the StatPearls question bank, assessing their outcomes, and providing insights into the integration of artificial intelligence (AI) technology in ophthalmology. Methods ChatGPT's performance was evaluated using 467 ophthalmology questions from the StatPearls question bank. These questions were stratified into 11 subcategories, four difficulty levels, and three generalized anatomical categories. The answer accuracy of GPT-3.5, GPT-4, and human participants was assessed. Statistical analysis was conducted via the Kolmogorov-Smirnov test for normality, one-way analysis of variance (ANOVA) for the statistical significance of GPT-3 versus GPT-4 versus human performance, and repeated unpaired two-sample t-tests to compare the means of two groups. Results GPT-4 outperformed both GPT-3.5 and human professionals on ophthalmology StatPearls questions, except in the "Lens and Cataract" category. The performance differences were statistically significant overall, with GPT-4 achieving higher accuracy (73.2%) compared to GPT-3.5 (55.5%, p-value < 0.001) and humans (58.3%, p-value < 0.001). There were variations in performance across difficulty levels (rated one to four), but GPT-4 consistently performed better than both GPT-3.5 and humans on level-two, -three, and -four questions. On questions of level-four difficulty, human performance significantly exceeded that of GPT-3.5 (p = 0.008). Conclusion The study's findings demonstrate GPT-4's significant performance improvements over GPT-3.5 and human professionals on StatPearls ophthalmology questions. Our results highlight the potential of advanced conversational AI systems to be utilized as important tools in the education and practice of medicine.